Application of Visual-Inertial SLAM for 3D Mapping of Underground Environments

António Ferreira, José Almeida and Eduardo Silva
INESC TEC - INESC Technology and Science
(formerly INESC Porto) and ISEP/IPP - School
of Engineering, Polytechnic Institute of Porto

***Abstract*—The underground scenarios are one of the most challenging environments for accurate and precise 3d mapping where hostile conditions like absence of Global Positioning Systems, extreme lighting variations and geometrically smooth surfaces may be expected. So far, the state-of-the-art methods in underground modelling remain restricted to environments in which pronounced geometric features are abundant. This limitation is a consequence of the scan matching algorithms used to solve the localization and registration problems.**

This paper contributes to the expansion of the modelling capabilities to structures characterized by uniform geometry and smooth surfaces, as is the case of road and train tunnels. To achieve that, we combine some state of the art techniques from mobile robotics, and propose a method for 6DOF platform positioning in such scenarios, that is latter used for the environment modelling.

A visual monocular Simultaneous Localization and Mapping (MonoSLAM) approach based on the Extended Kalman Filter (EKF), complemented by the introduction of inertial measurements in the prediction step, allows our system to localize himself over long distances, using exclusively sensors carried on board a mobile platform. By feeding the Extended Kalman Filter with inertial data we were able to overcome the major problem related with MonoSLAM implementations, known as scale factor ambiguity. Despite extreme lighting variations, reliable visual features were extracted through the SIFT algorithm, and inserted directly in the EKF mechanism according to the Inverse Depth Parametrization. Through the 1-Point RANSAC (Random Sample Consensus) wrong frame-to-frame feature matches were rejected.

The developed method was tested based on a dataset acquired inside a road tunnel and the navigation results compared with a ground truth obtained by post-processing a high grade Inertial Navigation System and L1/L2 RTK-GPS measurements acquired outside the tunnel. Results from the localization strategy are presented and analyzed.

I. Introduction

Over the last few years some successful underground mobile modelling implementations were documented [1] [2] [3]. These approaches, designed specifically to operate in mines, are characterized by one common aspect: they all use laser range finder sensors as the main (and in some cases the only) source of information. The model is built by placing laser range finder scans in a virtual three-dimensional world – process called registration. For this purpose, relative position and orientation between scans have to be determined. In previous approaches, this task is accomplished via a scan matching algorithm [7], which restricts the systems to non-uniform structures, since this technique requires that notorious and well-differentiated geometric features stand out along overlapping scans.

Our work extends the underground mobile modelling systems to galleries characterized by uniform and smooth surfaces. In this type of scenario the scan matching approaches are condemned to failure, so the previous state-of-the-art systems become ineffective. Without artificial landmarks and no access to Global Positioning Systems, self-localization becomes an hard problem. In inertial based localization the errors accumulated over time cause a monotonic growth in localization uncertainty. On the other hand, a vision based approach may be affected by the lighting conditions, additionally, the parametrization of landmarks far from the cameras raises extra difficulties due to the depth uncertainty.

Similarly to [3], our solution uses 2D laser range finders to gather a sequence of vertical scans along the gallery. Absolute position and orientation of each scan is computed by an independent localization process, that estimates the systems' trajectory based on inertial measurements and a sequence of images.

We employ an alternative localization solution to overcome both the structural monotony and the lack of Global Positioning Systems, adopting the SLAM (Simultaneous Localization and Mapping) concept [8] [9] to estimate the platforms localization in 6DoF (Six Degrees of Freedom). Following the traditional approach, the probabilistic SLAM algorithm is based on the EKF (Extended Kalman Filter). Since for landmarks far from the cameras, stereoscopic systems do not provide satisfactory depth measurements, a visual monocular algorithm was implemented instead, ensuring tracking of landmarks at any depth.

In order to identify visual landmarks to be used in the SLAM algorithm, highly distinctive visual features, invariant to scale, rotation and linear illumination variations, are extracted from the images using the SIFT algorithm [11]. To each feature is assigned at least one descriptor, that embodies the image properties in the features' neighborhood. The descriptors are used to establish the frame-to-frame feature matches.

Our system combines another advanced state-of-the-art methods such as Inverse Depth Parametrization [5], and the 1-Point RANSAC algorithm [6], for outlier rejection.

Through the Inverse Depth Parametrization, undelayed initialization of landmarks within the EKF framework be-

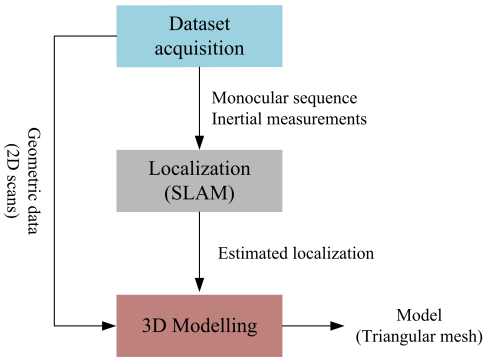

Fig. 1: High level system architecture

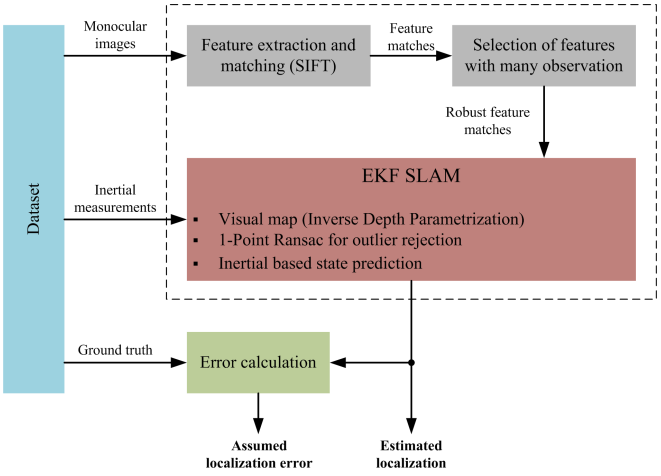

Fig. 2: Localization algorithm overview

comes possible. However another major problem of monocular SLAM applications still needs to be solved: a single camera moving through the scene does not provide metric measurements, leading to scale ambiguity in the estimated map and motion. As suggested in [4] inertial measurements, provided by a low-cost IMU, feed the filter with metric data in order to prevent the scale factor degeneration. This strategy keeps the map and motion estimates constrained to the meaningful metric system, in our case for distances over more than one hundred meters.

To build the model, all vertical cross sections are placed on a common reference frame according to the localization estimates, resulting in a point cloud model, which is finally converted into a triangular mesh through the Ball Pivoting Algorithm [10], to reach a more explicit representation without information losses. Texture captured by the cameras is also added to the model to enhance the visual realism.

This document is organized as follows: Section II presents a brief architecture description with emphasis on the localization and modelling algorithms. Section III is devoted to the dataset acquisition that takes places inside a road tunnel. We then present and discuss our implementation results (Section IV) and finally, Section V, provides a conclusion and sets some future goals.

II. System Architecture

Our system is divided in three main blocks, executed by the following order: data acquisition, localization and three-dimensional modelling (see Fig. 1).

In the first step, a sensor platform mounted on board a car is used to collect a wide range of synchronized measurements inside the underground galleries, including images captured by two CCD cameras, 2D scans from two laser range finders and inertial measurements provided by a low cost inertial measurement unit. The platform carries also a INS/GPS system that gives accurate ground truth information, used to measure the performance of our localization strategy.

The localization estimation and modelling tasks are preformed offline based on this data, according to the methods described next.

A. Localization Algorithm

In underground galleries it is expected to find reliable visual features that can be used as reference points to build the SLAM map. The process starts with a feature pre-selection stage (see Fig. 2) to fulfill the following objectives:

- Reduce the computational complexity of the SLAM cycle, by performing feature extraction and frame-to-frame matching in advance. The feature extraction is accomplished by the SIFT algorithm [11], that produces descriptors invariant to scale, orientation, and linear illumination changes, used to compute the frame-to-frame feature matches;
- Identify features with large number of observations and use only those to build the map. By doing so, we intend to minimize the computational demands, ensuring that all landmarks in the map persist over an acceptable frame interval.

1) State Vector: The SLAM cycle is implemented according to the EKF method. The state vector stores the localization and map states. Since the system does not have prior information about the environment, the initial state vector includes only 9 states related to the platforms' localization: position x^n, orientation Θ^n (expressed in terms of Euler angles) and velocity v^n, all defined in the local level reference frame (see Fig. 3).

$$x(k) = (x_b)^n(k) = \begin{bmatrix} x^n(k) \\ \Theta^n(k) \\ v^n(k) \end{bmatrix} \tag{1}$$

As new landmarks are observed, the state vector is expanded to accommodate the respective states (equation 2).

$$x(k) = \begin{bmatrix} (x_b)^n(k) \\ L_1(k) \\ L_2(k) \\ \vdots \\ L_n(k) \end{bmatrix} \tag{2}$$

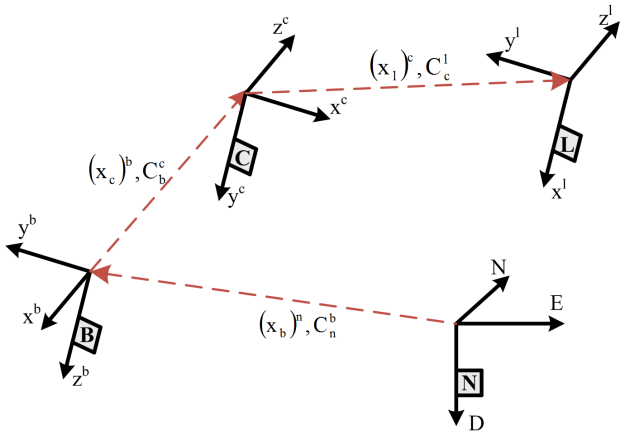

Fig. 3: Reference frames used in the localization and modelling algorithms. Local level frame (N), body frame (B), camera frame (C) and laser range finder frame (L).

Initially, each landmark L_i is coded in the SLAM map using the Inverse Depth Parametrization [5], which requires six parameters (Fig. 4): position of the cameras' optical center at the moment of first observation $[x_i^n\ y_i^n\ z_i^n]$, azimuth θ_i and elevation ϕ_i angles of the projection ray that passes through the optical center and the landmark, and finally the inverse of the distance ρ_i between the optical center and the landmark in the world (inverse depth).

$$L_i = [x_i^n,\ y_i^n,\ z_i^n,\ \theta_i,\ \phi_i,\ \rho_i]^T \tag{3}$$

The state uncertainty of this overparameterized representation can be modelled by Gaussian distributions, regardless to the distance between the landmark and the camera, therefore this is an efficient and accurate solution for undelayed initialization of new landmarks within the EKF. The EKF computational complexity grows quadratically with respect to the state vector dimension, so when the uncertainty in the landmark's location reveals a Gaussian behavior, indicated by the linearity index introduced in [12], the conversion to the standard Cartesian representation is accomplished applying the formula below:

$$\begin{bmatrix} L_{xi} \\ L_{yi} \\ L_{zi} \end{bmatrix} = \begin{bmatrix} x_i^n \\ y_i^n \\ z_i^n \end{bmatrix} + \frac{1}{\rho_i} m(\theta_i, \phi_i) \tag{4}$$

being $[L_{xi},\ L_{yi},\ L_{zi}]$ the Cartesian coordinates of the landmark and $m(\theta_i, \phi_i)$ a unitary vector (see Fig. 4), calculated from the azimuth and elevation angles:

$$m(\theta_i, \phi_i) = \begin{bmatrix} -cos(\phi_i)sin(\theta_i) \\ sin(\phi_i) \\ cos(\phi_i)cos(\theta_i) \end{bmatrix} \tag{5}$$

2) Landmark Initialization: From the six parameters that define an Inverse Depth landmark, only the azimuth and elevation angles need to be computed, since the camera position is already defined in the state vector, and the initial inverse depth consists on a fixed value defined in advance. To compute the angles, the feature is first projected from the image to the camera reference frame, using the pinhole camera model. A distortion model is applied next to compensate for the lens distortion. From this operation results a three-dimensional non-unitary vector h^c with the same orientation as the projection ray. The vector expressed in the navigation frame is given by:

$$h^n = C_b^n C_c^b h^c \tag{6}$$

where C_b^n and C_c^b are the rotations matrices from the body frame to the navigation frame and from the camera frame to the body frame, respectively (see Fig. 3).

From h^n, the orientation angles can be finally computed as follows:

$$\begin{bmatrix} \theta_i \\ \phi_i \end{bmatrix} = \begin{bmatrix} arctan(-h_x^n,\ h_z^n) \\ arctan\left(h_y^n,\ \sqrt{(h_x^n)^2 + (h_z^n)^2}\right) \end{bmatrix} \tag{7}$$

3) Landmark Prediction and Outliers Rejection: At the update step of the Extended Kalman Filter the position of the features observed in the image is compared to the expected projection of the map landmarks in the image. The projection of a landmark in the map to the image starts with the transformation from the navigation frame to the camera frame:

$$h^c = C_b^c C_n^b \left(\rho_i \left(\begin{bmatrix} x_i^n \\ y_i^n \\ z_i^n \end{bmatrix} - (x_b)^n - C_b^n (x_c)^b \right) + m(\theta_i, \phi_i) \right) \tag{8}$$

The distortion model is then applied to h^c, followed by the pinhole model, to determine the projection in the image.

Finally, wrong feature matches are rejected through the 1-Point RANSAC algorithm [6], that takes into account the prior probabilistic distributions maintained by the EKF to reduced the minimal sample size to only one feature match, significantly reducing the computational complexity associated with the standard RANSAC algorithm.

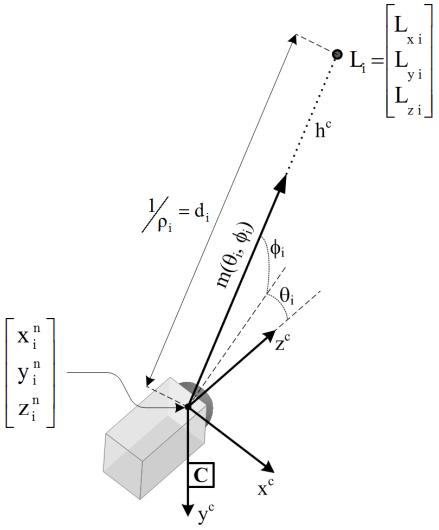

Fig. 4: Representation of the Inverse Depth parameters

4) Inertial Based State Prediction: To avoid the scale factor ambiguity, the main limitation of monocular SLAM caused by the absence of metric information, inertial measurements from a low cost IMU are injected in the EKF prediction step. Since the map landmarks are static, only the platform localization states are subjected to the motion model, that consists on the inertial mechanization in the local level reference frame, respecting the following equations:

$$\begin{bmatrix} x^n(k) \\ \Theta^n(k) \\ v^n(k) \end{bmatrix} = \begin{bmatrix} x^n(k-1) + v^n(k)\Delta t \\ \Theta^n(k-1) + E_b^n w^b(k)\Delta t \\ v^n(k-1) + \left(C_b^n a^b(k) + g^n\right)\Delta t \end{bmatrix} \quad (9)$$

where the IMU inputs are identified by a^b and w^b, respectively the linear accelerations and angular velocities, measured in the body reference frame. C_b^n is the direction cosine matrix obtained from the platform orientation and E_b^n is a 3 by 3 matrix that converts the angular velocities into the Euler angles rate of change:

$$E_b^n = \begin{bmatrix} 1 & sin(\phi)tan(\theta) & cos(\phi)tan(\theta) \\ 0 & cos(\phi) & -sin(\phi) \\ 0 & sin(\phi)sec(\theta) & cos(\phi)sec(\theta) \end{bmatrix} \quad (10)$$

B. Modelling Algorithm

The three-dimensional model is constructed by placing all gallery cross-sections, taken by the vertical laser range finder, into a common coordinate system.

First, laser range finder scans, initially expressed in polar coordinates, are converted to the Cartesian coordinate system with origin matching the center of the laser range finder. Next, specific position and orientation of each scan is derived from the two closest localization points in time. Given the calibration parameters that describe the spatial relationship between sensors, determined in advance, and using the calculated scan localization, all vertical cross-sections are transformed to the local level frame according to the formula below:

$$P^n = C_b^n \left(C_c^b \left(C_l^c \left(P^l - (x_l)^c \right) - (x_c)^b \right) - (x_b)^n \right) \quad (11)$$

where P^n is the final point in the local level frame, whereas P^l refers to the original point in the sensor Cartesian system. The rotation matrices C_c^b and C_l^c establish the rotation from camera to body and laser to camera reference frames, respectively (see Fig. 3). Whereas $(x_c)^b$ define the camera position in the body frame and $(x_l)^c$ the laser position with respect to the camera frame. Finally C_b^n and $(x_b)^n$ enclose the rigid body transformation from the body to the local level reference frame.

After applying formula (11) to all points of all scans, a point cloud model is achieved. Usually, the interpretation of point clouds is not easy due to lack of surfaces. To improve the scene's perception, original surfaces are reconstructed by converting the point cloud into a triangular mesh, using the Ball Pivoting Algorithm (BPA) [10]. The models realism is also enhanced by introducing texture information captured by the cameras.

To reduce the noise and produce smoother surfaces, a Laplacian filter is applied to the whole triangular mesh, computing a new position for each vertex according to local information given by adjacent points.

Both the point cloud model and the triangular mesh are coded in the VRML format to be displayed in a virtual reality application.

III. Dataset Acquisition

Solving the localization and modelling problems demands previous acquisition of a variety of measurements. To this purpose different types of sensors where assembled in a rigid platform (see Fig. 5), which in turn is mounted on top of a car.

The vertical cross-sections are taken by the vertical laser range finder (SICK LMS-200) at 75Hz with an angular resolution of 1°. There are two pointing-forward cameras (JAI CB-080GE), arranged in a stereoscopic configuration, with a resolution of 1032(h)x778(v) and controlled by an external trigger at a frame rate of 7 fps. Only the images from the left camera are used in our SLAM system.

The low cost IMU (MicroStrain 3DM-GX1), placed above the left camera, gives the linear acceleration and angular velocity measurements used in the EKF prediction step, at a frequency of 100Hz.

Ground truth with a 400 Hz rate is obtained by a tactical grade INS/GPS system (iMAR iNAV-FMS-E) placed in the center of the platform. This system provides raw inertial data and GPS measurements acquired outside the gallery, that are post-processed in a commercial software (Waypoint Inertial Explorer) to produce an accurate trajectory estimation. This trajectory is only used as ground truth to evaluate the SLAM performance.

All system reference clocks are synchronized with respect to GPS clock, to assure a consistent time base.

The data acquisition experiment took place on a road tunnel with approximately 140 meters located at Vilar de Luz – Porto (see Fig. 6). All data were correctly logged. However the

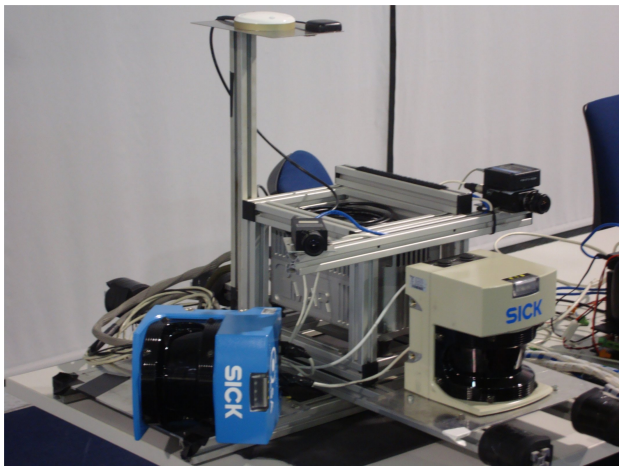

Fig. 5: Sensor platform used for data acquisition

Fig. 6: Preparation for the data acquisition experiment in the tunnel area

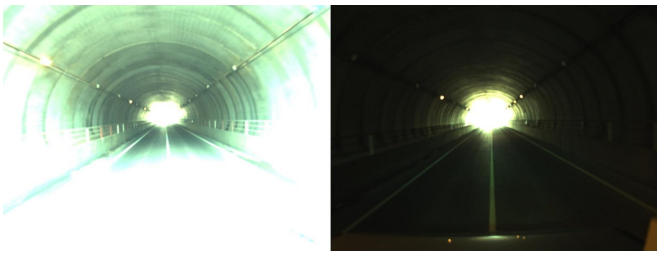

Fig. 7: Image instability as consequence of the illumination variations along the tunnel.

images reflect the huge lighting variations between the interior and exterior of the tunnel (see Fig. 7).

IV. Results

An accurate localization estimate is crucial to obtain a reliable model reproducing the real gallery characteristics. Using the ground truth trajectory the error associated with the estimated localization is determined. Furthermore, to realize the benefits of fusing inertial and visual measurements, both inertial navigation and MonoSLAM approaches were implemented, and the results are compared with the ones achieved by the inertial and visual SLAM approach.

The errors in the position states for each method are outlined in Fig. 8. The path calculated by MonoSLAM shows the worst results due to the scale ambiguity, accumulating an error of 11.7 meters. As expected, inertial navigation drifts with time due to error integration, resulting in a total drift of 8.7 meters. Our approach produces the smallest error, showing the advantage of inertial and visual data fusion, with a maximum value of 1.29 meters and an error of 0.95 meters at the final position. The insertion of inertial measurements in the MonoSLAM mechanism successfully prevents the scale factor ambiguity, whereas visual data contributes to the inertial drift compensation, particularly to the orientation states correction.

The distribution of the features along the image is shown in Fig. 9. The image space is equally divided in four quadrants and an histogram is computed for each portion. It can be seen that the top quadrants provide the most reliable features, in terms of number and long observation sequence. The landmarks introduced in the SLAM map are observed over more

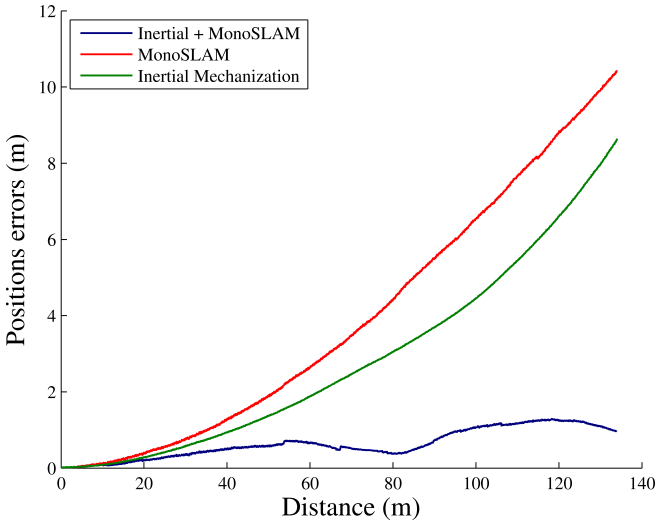

Fig. 8: Position errors produced by each localization strategy: SLAM fusing inertial and visual data (blue line), inertial mechanization (green line) and monocular SLAM (red line)

then 6 frames, and the most persistent ones last a maximum of 50 frames. The short periods of observation reduce the possibility of observing sufficient parallax to convert inverse depth landmarks to the Cartesian form (see Fig. 10).

As the system approaches the end of the tunnel, the effects of image saturation are visible at the final moments in Fig. 10. In the last 20 meters, the 1-Point RANSAC algorithm rejects a considerable amount of wrong feature matches, and the respective landmarks are deleted from the SLAM map.

As previously mentioned, the point cloud models can become really hard to interpret, depending on the view point

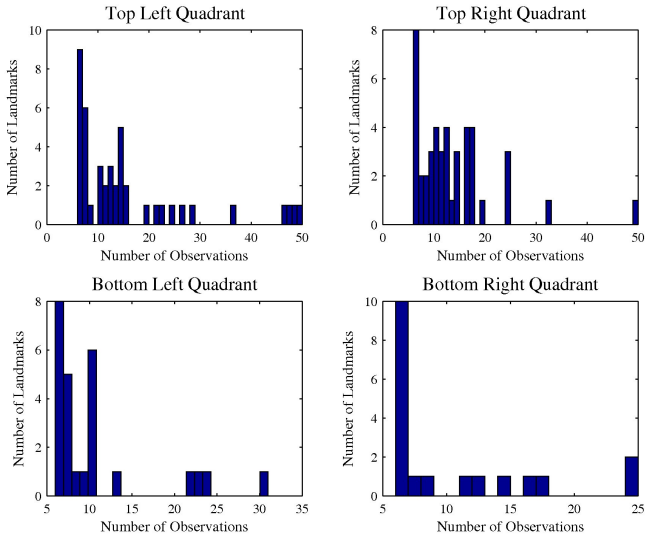

Fig. 9: Histograms that represent the total number of features in respect to the number of observations, for image sub-regions of equal size.

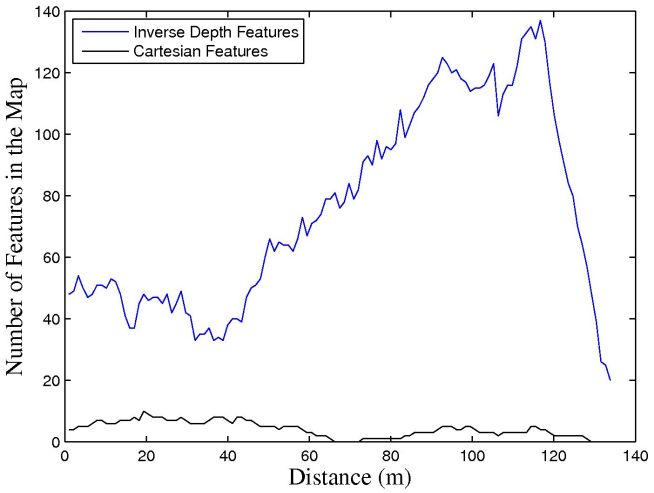

Fig. 10: Landmarks parametrization in the SLAM map along the trajectory.

and scale. In order to reach a more explicit and realistic representation, a triangular mesh is constructed from the point cloud without data losses, through the Ball Pivoting Algorithm. In the final step the surfaces are filtered by a Laplacian smoother, and texture acquired by the cameras is added to the model (see Fig. 11).

V. Conclusion

The development of a mobile modelling system for large scale underground environments raises some difficult challenges, especially when dealing with monotonous geometry. Based on inertial and visual data we have implemented a localization method that does not depend on the geometric properties of the environment, thus it is specifically suited to operate inside smooth shape galleries like traffic tunnels.

Through localization results the benefit of fusing inertial data within the MonoSLAM strategy became evident. In the most aggressive configuration, with a pointing forward camera, forward motion and large illumination variance, our localization estimate reached an error of 0.95% of the total displacement, which constitutes a quite impressive accomplishment given the low cost sensors used.

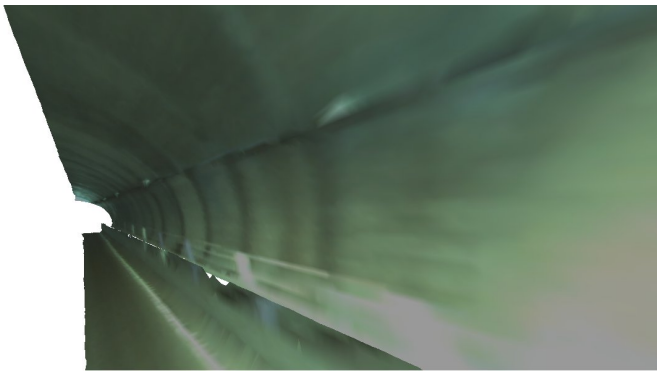

Fig. 11: Triangular mesh model after Laplacian filtering.

Despite the poor image quality, reliable visual features and descriptors where extracted by the SIFT algorithm, exploiting the algorithm's immunity to rotation scale and linear illumination variations, enabling robust frame-to-frame feature matching.

In the future, localization accuracy could be improved by adding other types of information, for instance, laser range finder measurements to provide a better approximation of the landmarks initial depth. A stereo vision system will also be implemented to enable instant computation of close landmark coordinates. The use of cameras with larger field of view will also be beneficial, enabling the observation of landmarks with high parallax and hence low depth uncertainty.

Acknowledgment

This work was supported by QREN – Project n° 7865 "FlexiMap3D".

References

[1] Andreas Nüchter, Hartmut Surmann, Kai Lingemann, Joachim Hertzberg and Sebastian Thrun, *6D SLAM with an Application in Autonomous Mine Mapping*, IEEE International Conference on Robotics and Automation (ICRA), pages 1998–2003, 2004.

[2] Daniel Huber and Nicolas Vandapel, *Automatic Three-dimensional Underground Mine Mapping*, The International Journal of Robotics Research, volume 25, pages 7–17, January 2006.

[3] Sebastian Thrun, Dirk Hähnel, David Ferguson, Michael Montemerlo, Rudolph Triebel, Wolfram Burgard, Christopher Baker, Zachary Omohundro, Scott Thayer and William Whittaker, *A System for Volumetric Robotic Mapping of Abandoned Mines*.

[4] Pedro Pinies, Todd Lupton, Salah Sukkarieh, Juan D. Tardós, *Inertial Aiding of Inverse Depth SLAM Using a Monocular Camera*, IEEE International Conference on Robotics and Automation (ICRA), pages 2797–2802, 2007.

[5] Javier Civera, Andrew J. Davison and J. M. M. Montiel, *Inverse Depth Parametrization for Monocular SLAM*, IEEE Transactions on Robotics, volume 24, number 5, pages 932–945, October 2008.

[6] Javier Civera, O. Garcia, Andrew J. Davison and J. M. M. Montiel, *1-Point RANSAC for EKF-Based Structure from Motion*, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), October 2009.

[7] Paul J. Besl and Neil D. McKay, *A Method for Registration of 3-D Shapes*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 14(2), pages 239–256, 1992.

[8] R. Smith, M. Self and P. Cheeseman, Estimating Uncertain Spatial Relationships in Robotics, In I. Cox and G. Wilfong, editors, *Autonomous Robot Vehicles*, Springer-Verlab, 1990.

[9] J. J. Leonard and Durrant H. Whyte, *Simultaneous Map Building and Localization for an Autonomous Mobile Robot*, IEEE International Conference on Intelligent Robots and Systems (IROS), Osaka, Japan, 1991.

[10] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin, *The Ball-Pivoting Algorithm for Surface Reconstruction*, IEEE Transactions on Visualization and Computer Graphics (TVCG), 5(4), pages 349–359, 1999.

[11] David G. Lowe, *Distinctive Image Features from Scale-Invariant Keypoints*, International Journal of Computer Vision (IJCV), 60(2), pages 91–110, 2004.

[12] Javier Civera, Andrew J. Davison, J. M. M. Montiel, *Inverse Depth to Depth Conversion for Monocular SLAM*, International Conference on Robotics and Automation (ICRA), pages 2778–2783, 2007.